CSCI 1850: Deep Learning in Genomics  
Spring 2021

http://cs.brown.edu/courses/csci1850

Mar 23, 2021  
Tuesday

Instructor: Ritambhara Singh  
Format: Online (Synchronous)  
Time: TTh 10:30-11:50 AM
Today’s outline

• Denoising single-cell data using autoencoders: DCA

Section IV: Interpretation of deep learning models

• CNN filter weights and perturbation analysis
  o DeepBind
  o DeepSEA

• Class Activity

• DeepMotif Dashboard: Visualization methods for TFBS predictions
  o Saliency Map
DCA: Input

Gene expression heatmap showing the expression levels across different cells.
DCA: Model

6 genes

Cells

Genes

Expression
Low High

Bottleneck layer

Denoised output

Encoder Decoder
DCA: Output

Represented by the noise model

Replace MSE loss with ZINB loss
What is ZINB?

Zero inflated negative binomial distribution
What is ZINB? How do we model count data?

- Poisson distribution: Mean = Variance
- Negative Binomial: Mean ≠ Variance

RNA-Seq data can have variable variance due to noise (artifacts, unknown causes)

Image courtesy: https://www.vosesoftware.com/riskwiki/NegativeBinomial.php
What is ZINB?

Zero inflated negative binomial distribution

Why zero inflated version for single-cell?

Image courtesy: https://stats.stackexchange.com/questions/264528/is-this-zero-inflated-negative-binomial-distribution
DCA: Summary

\[ \text{ZINB} \left( x; \pi, \mu, \theta \right) = \pi \delta_0 \left( x \right) + (1 - \pi) \text{NB} \left( x; \mu, \theta \right) \]

\[ \hat{\Pi}, \hat{M}, \hat{\Theta} = \arg\min_{\Pi, M, \Theta} \text{NLL}_{\text{ZINB}} \left( x; \Pi, M, \Theta \right) \]

 Represents the noise model

[Diagram: Input x \rightarrow ZINB (x | \mu, \theta, \pi) \rightarrow Output \rightarrow Denoised output]

Cells

Genes

Expression

Low

High

Encoder

Decoder

Dropout \( \pi \)

Dispersion \( \theta \)

Mean \( \mu \)
DCA: Results

Simulations

Muddy point: How are simulations created?
Questions?
In which paper/work did we see interpretation using linear regression?
DeepBind and interpretation : Model

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning
DeepBind and interpretation: Filter weights
Perturbation analysis

Model training

ATATCGCTCAGCTCGATCG

Binding site?
Perturbation analysis

ACATCGCTCAGCTCGATCG

Trained model
Perturbation analysis

ATATCGCTCTGCTCGATCG

Binding site?

Trained model
DeepBind and interpretation: Perturbation analysis

Mutation Maps

What information can we learn from these?

- Familial hypercholesterolemia
- Cancer risk variant
DeepSEA and interpretation: Model

Predicting effects of noncoding variants with deep learning–based sequence model
DeepSEA and interpretation: Model

Predicting effects of noncoding variants with deep learning–based sequence model.

Output:
- variant functionality prediction
- predicted chromatin effect
- predicted allele-specific chromatin profile

Input:
- genomic sequences (1,000 bp)

Training data:
- ENCODE
- Roadmap Epigenomics chromatin profiles

Deep convolutional network (DeepSEA)
DeepSEA and interpretation: Perturbation analysis

*In silico* mutagenesis effects on TF binding

Score = log fold change

What can be the potential drawback of perturbation analysis?
Questions?
Backpropagation is back!

Applying back-propagation

Calculating gradient

\[ g(f(k(i))) \]

\[ \text{Error} = \frac{1}{2} (\text{predicted} - \text{actual})^2 = \frac{1}{2} (f(k(i)) - 1)^2 \]

\[ f(k(i)) = f(h) = h_1w_5 + h_2w_6 \]

\[ k(i) = h_1 = i_1w_1 + i_2w_2 \]

\[ \frac{d(g(f(k(i))))}{d(w_1)} = \frac{d(\text{Error})}{d(w_1)} = \frac{d(\text{Error})}{d(f(h))} \cdot \frac{d(f(h))}{d(h_1)} \cdot \frac{d(h_1)}{d(w_1)} \]

Chain rule

What do we need to change to calculate this gradient w.r.t to \( i_1 \)?
Applying back-propagation

Calculating gradient w.r.t input $i_1$

$$g(f(k(i)))$$

Error $= \frac{1}{2} (predicted - actual)^2 = \frac{1}{2} (f(k(i)) - 1)^2$

$$f(k(i)) = f(h) = h_1w_5 + h_2w_6$$

$$k(i) = h_1 = i_1w_1 + i_2w_2$$

$$k(i) = h_2 = i_1w_3 + i_2w_4$$

$$\frac{d (g(f(k(i))))}{d(i_1)} = \frac{d(\text{Error})}{d(i_1)} = \frac{d(\text{Error})}{d(f(h))} \cdot \frac{d(f(h))}{d(h_1)} \cdot \frac{d(h_1)}{d(i_1)} + \frac{d(\text{Error})}{d(f(h))} \cdot \frac{d(f(h))}{d(h_2)} \cdot \frac{d(h_2)}{d(i_1)}$$

Chain rule
Class Activity (Think-pair-share) [10 mins]

Calculate gradient w.r.t $i_2$

$g(f(k(i)))$

$\text{Error} = \frac{1}{2} \left( \text{predicted} - \text{actual} \right)^2 = \frac{1}{2} \left( f(k(i)) - 1 \right)^2$

$f(k(i)) = f(h) = h_1w_5 + h_2w_6$

$k(i) = h_1 = i_1w_1 + i_2w_2$

$k(i) = h_2 = i_1w_3 + i_2w_4$

https://docs.google.com/document/d/11kMROd4JBu6iqNoRS_lCx0X298Fe6FsWqO2CRlQ-wDM/edit?usp=sharing

Questions?
Visualization methods for TFBS predictions


DEEP MOTIF DASHBOARD: VISUALIZING AND UNDERSTANDING GENOMIC SEQUENCES USING DEEP NEURAL NETWORKS

JACK LANCHANTIN, RITAMBHARA SINGH, BEILUN WANG and YANJUN QI
DeepMotif Dashboard Setup

1. Convolutional (CNN)  
   (short local patterns, or motifs)

2. Recurrent (RNN)  
   (long term dependencies)

3. Convolutional-Recurrent (CNN-RNN)  
   (long term dependencies among motifs)

Where have we seen this model?
DeepMotif Dashboard Visualization methods

- Saliency Maps
- Temporal outputs (next class)
- Class based optimization (next class)
Saliency map

Which nucleotides are most important for classification?
Saliency map

\[
S_+(X) \approx w^T X + b = \sum_{i=1}^{\|X\|} w_i x_i
\]
Saliency map

\[ S_+(X) \approx w^T X + b = \sum_{i=1}^{\mid X \mid} w_i x_i \]

\[ w = \frac{\partial S_+}{\partial X} \bigg|_{x_0} = \text{“saliency map”} \]
Saliency map

This movie has one of the best plots I have seen

Positive sentiment

This movie has one of the best plots I have seen

= important for classification

Image courtesy: https://pdfs.semanticscholar.org/b82d/2edf99303589ec40e87a91e4d08470baf743.pdf
Saliency map

\[ X \rightarrow S_+ \]

positive binding site

<table>
<thead>
<tr>
<th>Positive Test Sequence</th>
<th>TGTC6CATCCTATTGGCCACGTTAGTCACATGGCCCCACCTGGCTGCAAAAGCACTGGGAAAC6TAGCTTTCTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saliency Map</td>
<td>![Saliency Map Image]</td>
</tr>
</tbody>
</table>

\[ = \text{important nucleotide for prediction} \]

Image courtesy: https://pdfs.semanticscholar.org/b82d/2edf99303589ec40e87a91e4d08470baf743.pdf
Questions?
Upcoming

Course website: [http://cs.brown.edu/courses/csci1850](http://cs.brown.edu/courses/csci1850)

Section IV: Interpretation of deep learning models

- **March 25**: Temporal outputs of RNNs, Class based optimization, Attention

- Course project deadlines **tonight at 11:59PM**
  - Final changes to your code
  - Final submission to Kaggle
  - Project report (submission via Canvas)

- Homework 3 released, due **April 01 at 11:59PM**
Wrap up

What was the clearest point today?

What was the muddiest point today?

https://forms.gle/GSUYj1XujKvpfJYQA